Issues in Informing Science and Information Technology

Automatically Grading Essays with Markit©

Robert Williams and Heinz Dreher
Curtin University of Technology, Perth, WA, Australia

williamsr@cbs.curtin.edu.au dreherh@cbs.curtin.edu.au

Abstract

Markit© is an Automated Essay Grading (AEG) system capable of running on typical desktop PC platforms. Its performance compares favourably with human graders and with commercially available systems. A distinct advantage of Markit over existing commercial systems is that it requires only one model answer against which the student essays are compared. In this paper we report on a trial of Markit with second year law students' essays.

Keywords: Automated Essay Grading (AEG); Natural Language Processing (NLP); semantics; electronic thesaurus.

Introduction

At IS2002 in Cork, Ireland (Palmer et al. 2002), we reported on a trial of a commercially available computer grading system as used to automatically grade first year university student essays. The results were encouraging, but we felt confident that certain perceived limitations could be overcome by applying ourselves to building our own system. Since then we have developed our own prototype and are investigating its performance with a wide range of subject areas and year levels. We have named our system Markit©.

In order to automate the grading of essays some method of capturing the meaning of the words, sentences and paragraphs must be found. Representing the semantics of a text document for computational uses is one method but is problematic. How can we formally code the meanings of words, phrases, sentences, paragraphs and so on, so that useful computational work can be done robustly, effectively and efficiently? Such semantic representation and computational work has been applied to the problem of essay grading in an endeavour to overcome some of the limitations we discussed with alternate approaches adopted by existing essay grading systems, but it may also be applied to related problems of text understanding, question answering, and qualitative feedback on assignments, for example.

In this article we present our work in developing the Markit system, now in prototype form and being used for Automated Essay Grading (AEG) and related applications.

Problems Identified from Previous Experience

The system we previously trialed, required us to manually grade 200 essays which were used to build a reference database against which candidate essays would be benchmarked and an appropriate score assigned. Cleary, if it were possible to use just one model answer to produce a reliable automated scoring system, it would be feasible to employ the system in the grading of tens and hun-

dreds rather than thousands of essays. Such a system would be more widely applicable. Markit uses just one model answer. The content of the model answer is determined by the instructor, and thus he/she has complete control over the knowledge against which the student essays are to be assessed. The breadth and depth of the knowledge required for very good answers can thus be determined by the instructor.

The cost of using the previous system was prohibitive for all but very large (in excess of 2000) essay numbers. Comparing costs can be an arbitrary exercise, and whilst the previously trialed system was approximately double that of manual grading, costs associated with Markit will need to be kept well below the cost of human grading if it is to come into popular usage. At this stage, we may be flexible in assigning costs to essay grading via Markit, preferring to acquire some experience with its use in a wide variety of circumstances. There are real costs however in doing any grading work, and these must be covered on a case by case basis. The costs which may be associated with product development will need to be recouped over a longer term when we are satisfied that Markit is performing at its optimum.

An important aspect of assessment relates to elapsed time taken to complete a job. Invariably, the requirement is for graders, human or otherwise, to produce results within a few days. We found that the elapsed time related to job submission and results provision can be longish, and figured that the availability of an in-house system would permit better control over this factor. Naturally, there would be many ways to control the elapsed-time variable, but having a system at one's fingertips is reassuring at the very least, and permits re-runs, and other experimentation to occur. From this standpoint, Markit suits our purposes very well indeed; we have control of the entire process from assignment and model answer submission through to results notification, meaning that security for example can be managed directly and effectively.

In our previous work we noted another serious limitation to an essay grading system – it can only grade by making a comparison with a given set of subject material. The model answer contains only a set body of knowledge and would grade the student on the part of that knowledge the student was able to demonstrate. Under such circumstances a 'brilliant' answer or essay, for example which drew on associations with material not part of the model answer, would score poorly. This problem, whilst recognized, has not been overcome in Markit's design. We will need to understand that Markit's grades are recommendations to human examiners and should be reviewed for appropriateness much as human scored essays required moderation and review in a selection of cases.

Probably the most advantageous aspect of Markit is its reliance on generally available technology as compared with specialized computing platforms needed to run a previously trialed system. Markit will operate on a standard Windows PC, and it will be developed in such a way that some of the system's components can be shrink-wrapped for widespread distribution and local use. This has been made possible due to the computational algorithm at the heart of Markit's design.

The Markit© System

Some readers will naturally be eager to discover the precise design of Markit's algorithms, but will also understand the proprietary nature of potentially commercially viable systems and thus the confidential nature of those designs. We are able however to characterize Markit's general design, although we look forward to satisfying readers' curiosity by performance data as we progress and expand our use across a wide variety of cases.

The Markit system relies on building a propriety representation of the knowledge contained in the model answer. A student essay is processed using a combination of NLP (Natural Language Processing) techniques to build the corresponding propriety knowledge representation. Pattern matching techniques are then employed to ascertain the proportion of the model answer knowl-

edge that is present in the student answer, and a grade assigned accordingly. An electronic version of Roget's Thesaurus is used to extract lexical information for the building of the document knowledge representation.

The technique allows a formal representation of free unseen text to be quickly and robustly built for further analysis by the Markit system. The approach used has a need for a semantic representation that does not need substantial hand coding of knowledge structures prior to use, and that can deal with unlimited unseen text. Many Natural Language Processing (NLP) systems use some kind of a parser to initially extract the syntax of sentences in a document as an initial step prior to further processing. Semantic analysis then follows. The use of Context Free Phrase Structure Grammar (CFPSG) parsers is commonly suggested in the literature. However CFPSG parsing cannot be used in all but simple toy domains. The reason for this is that free unseen text is very hard to parse, because the set of grammar rules required is very large, and the time taken to evaluate every possible parse tree generated is too great for a practical system. So while CFPSG parsing has been tried with the prototype system described in this article, it has been abandoned in favour of using "Chunking" to determine the phrases and clauses used for further processing. "Chunking" enables one to use grammar heuristics to derive noun phrases and verb clauses very quickly from unseen text. The problem of unrealistic parsing time is thus eliminated.

Markit's grading system is capable of producing perfect scores when grading a document against itself, the desirability of which can be appreciated.

Markit is at the prototype stage, and is undergoing further development. The system currently has 8 subsystems written in C++, Java and Visual Basic for Applications.

The extraction of information from Roget's Thesaurus (Roget, 1991) is slow, due to the fact that approximately 500 pages of a Microsoft Word document have to be scanned for each word in a sentence, using Visual Basic for Applications code. This process can take up to about 10 minutes for a 40 word sentence and clearly needs to be modified to access a database version of Roget's Thesaurus or equivalent. We are currently in negotiations with a publisher to have research access to a commercially available electronic dictionary/thesaurus. We will incorporate this in a database table, facilitating direct file access to each word in the thesaurus. Processing time will then be reduced to matter of seconds.

System performance otherwise is very good, with the non–Roget's Thesaurus related work taking only a few seconds for a 2 page document on a 1.9 Ghz Pentium 4 processor.

Markit Performance - the Law 252 Trial

Markit was trialed with student essays from a second year university law unit taught at the Curtin Business School (Willesee, 2003).

A lecturer in a School of Business Law kindly volunteered to assist by providing 66 essays from his unit Law 252 for Markit to assess. He also provided a model answer. This model answer was then processed to produce the model answer numerical summary. A selection of 20 student essays were then processed and graded against the model answer. Comparisons were made between the human scores and the automated scores from Markit. The results are shown in Table 1and Figure 1.

Table 1 - Human and Markit scores for Law 252 essays

Essay	Human	Markit	Absolute Difference	Human vs Markit
1	25	40	15	0.79
2	35	39	4	
3	45	29	16	
4	45	49	4	
5	50	63	13	
6	50	64	14	
7	50	68	18	
8	55	60	5	
9	60	59	1	
10	60	70	10	
11	70	60	10	
12	70	60	10	
13	70	64	6	
14	70	78	8	
15	75	66	9	
16	75	78	3	
17	80	71	9	
18	80	74	6	
19	80	81	1	
20	90	74	16	
Average	**61.75**	**62.35**	**8.90**	

From Table 1 (essay 1) it can be seen that a student essay was graded 25/100 by a human and 40 by Markit, resulting in a difference of 15 percentage points between the two. Now this student would presumably be very happy with such a score since it is higher than that of the Human grader. However the student author of essay 3 suffers from about the same error in absolute terms, but surely would be grossly dissatisfied with the failing grade. These extreme cases will need to be investigated with a view to understanding the reason. We have found in one much more extreme case that the error was due to the human grader rather than Markit. Obviously,

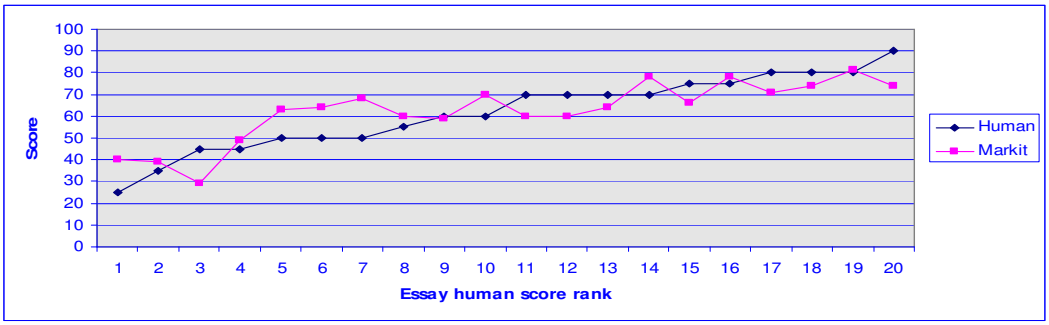

Figure 1 - Human versus Markit scores in ascending order of human scores for Law 252 essays

many more studies need to be conducted across time, subject matter, year levels, and so on, before we can appreciate the true worth of Markit's contribution. So, whilst the average differences between Human grades and Markit is acceptable, it is the large individual differences, particularly where the Markit grade is lower which need investigation and analysis.

Note the difference in the average marks is 0.60 %, and this is not significant with a p-value of 0.80 for a 2 tailed test of significance. The average error is 8.90%, and the Pearson correlation between the human and computer scores is 0.79, significant at the 1% level with a 2 tail test. This correlation is computed on the two scores for the same essay.

When we rearrange the data in ascending order of Markit scores, Figure 2 shows the trend. It appears that Markit is assigning scores using a moving average of the three neighbouring human scores. This is a highly desirable outcome, indicating that it is capturing the essence of the human grader's criteria, but of course Markit does not have access to the human scores. This characteristic has not yet been analysed statistically.

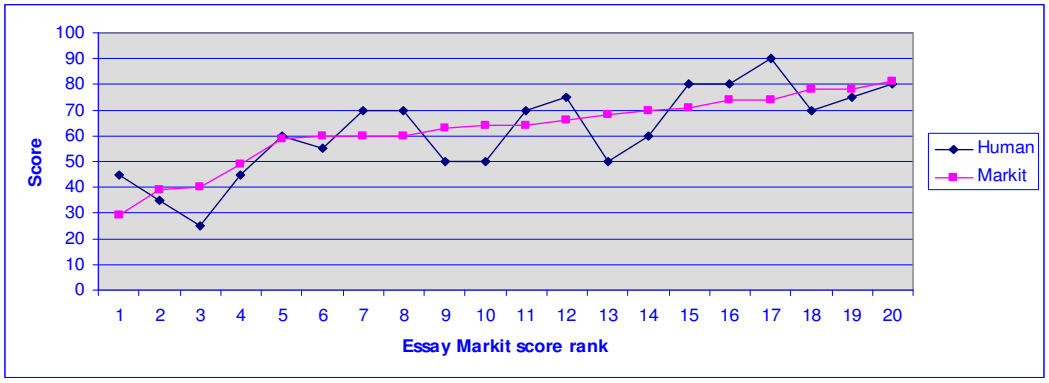

Figure 2 - Human versus Markit scores in ascending order of Markit scores for Law 252 essays

Markit Performance – compared to the IEA

As already mentioned, in 2001 we conducted a trial of a commercially available automated essay grading system (the IEA – Intelligent Essay Assessor, see Landauer et al. 1998) using essays from a first year Curtin University unit, Information Systems 100.

Nine of these essays were also graded by Markit. The top graded essay by the IEA gained 99%. This essay was then used as the model answer against which the others were compared using Markit. Table 2 and Figure 3 below show the results.

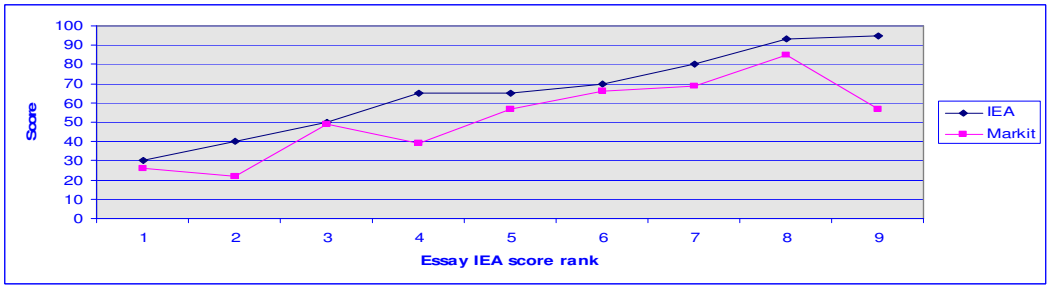

Figure 3 - IEA versus Markit scores in ascending order of IEA scores for IS 100 essays

Table 2 - IEA and Markit scores for IS 100 essays

Essay	IEA	Markit	Absolute Difference	IEA vs Markit
1	30	26	4	0.84
2	40	22	18	
3	50	49	1	
4	65	39	26	
5	65	57	8	
6	70	66	4	
7	80	69	11	
8	93	85	8	
9	95	57	38	
Average	**65.33**	**52.22**	**13.11**	

Note the difference in the average marks is 13.11 %, and this is significant with a p-value of 0.01 for a 2 tailed test of significance. The average error is 13.11%, and the Pearson correlation between the human and computer scores is 0.84, significant at the 1% level with a 2 tail test. This correlation is computed on the two scores for the same essay.

When we rearrange the data in ascending order of Markit scores, Figure 4 shows the trend. In this case, Markit appears to be assigning scores on a downward shifted moving average of the three neighbouring IEA scores, again not analysed statistically. This is not an ideal situation, but an adjustment factor could be built into Markit to shift the scores up. Further testing of Markit will determine the extra tuning parameters we may need to add to the scoring algorithm. Again, Markit does not have access to the IEA scores.

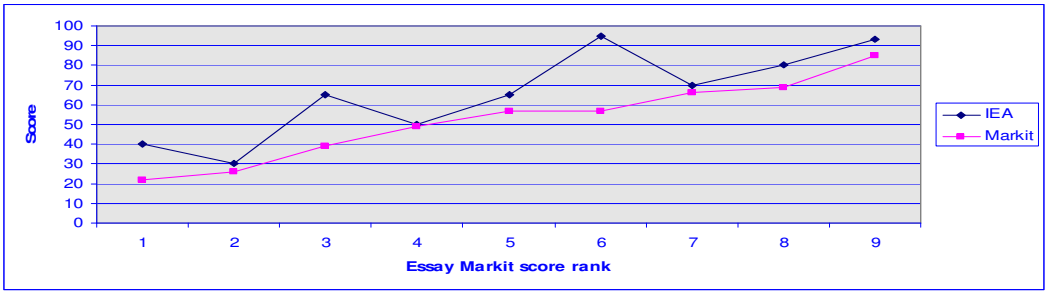

Figure 4 - IEA versus Markit scores in ascending order of Markit scores for IS 100 essays

Markit grades well when compared to the human marker. It tracks the human scores well, with an acceptable error rate. However it appears that we have tuned Markit to the Law 252 essays, as it scores lower than the IEA on the IS 100 essays, which has an average error of 13.11%. The Pearson correlations in both cases are acceptable, although we would like to see them at about 0.90. The IEA essays were not directly graded against a human marker, and the lower scores from Markit could be related to this.

The result of Markit's automated grading, when compared to human markers, is at the high end of published results for other AEG systems (Williams, 2001). Dessus et al., (2000) report that the highest correlations are found between human graders and Latent Semantic Analysis (LSA) based

techniques and are 0.80 and 0.86. In a trial of Markit we obtained correlations of 0.79 between the human marker and Markit, and, on a different set of essays, 0.84 between the Intelligent Essay Assessor grading and Markit grading.

Concluding Observations

The semantic representation as mentioned above lends itself to speedy processing by the grading subsystem. Comparisons between documents can then be made by looking for similar content, even if the documents use completely different wording. The programming of such content matching algorithms is relatively straightforward as a result of thesaurus based numerical array structure representations as used by Markit for comparing student essays against model answers.

Markit's performance is equally as good as other systems documented in the literature and yet the performance is achieved with minimal human grader input – only one model answer is required in comparison with some other systems which require several hundred human graded essays.

References

Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-text assessment in a virtual campus. *Proceedings CAPS'2000*. Paris : Europia, 13 –14 Dec.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis, *Discourse Processes*, *25*, 259-284.

Palmer, J., Williams, R., & Dreher, H. (2002). Automated essay grading system applied to a first year university subject – How can we do it better? *2002 Informing Science & IT Education Joint Conference: InSITE*, University College Cork, Ireland.

Roget, P. M. (1991). *Roget's thesaurus.* Project Gutenberg, http://promo.net/pg/

Williams, R. F. (2001). Automated essay grading: An evaluation of four conceptual models. In M. Kulski & A. Herrmann (Eds.), *New horizons in university teaching and learning: Responding to change.* Curtin University of Technology, Perth, Western Australia

Willesee, W. (2003). Can a computer grade an Internet uploaded law essay? *5th Conference on Computerisation of Law via the Internet*, Sydney, Australia. November.

Biography

Robert Williams has over 25 year's experience in the Information Systems industry, as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design, and programming, on a variety of mainframe, mini and personal computers, and a variety of operating systems and programming languages. Applications he has worked with include mathematical, statistical, bridge and road engineering, financial, corporate resource allocation, business simulation and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading systems. In 2001 he led a team of researchers in the School of Information Systems at Curtin University of Technology which conducted what is believed to be the first trial in Australia of an Automated Essay Grading system. Robert holds a Bachelor of Arts degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.

Heinz Dreher has been working in the Information Technology Systems domain for 33 years. His first position was as computer programmer. This was followed with a move into the tertiary education sector in 1972 as senior tutor in Electronic Data Processing (EDP). Currently he is on

sabbatical at the Institute of Interactive Computer Multimedia at the University of Technology Graz, Austria. His substantive position is in the School of Information Systems at Curtin University of Technology, Perth, Western Australia. Dr Dreher has expertise in Hypertext/Hypermedia systems and textual-knowledge-based systems, Computer Supported Co-operative Work (CSCW), Computer Mediated Communications (CMC), Project Management, Prototyping systems, Human Problem Solving Strategies, Decision Support Technologies, Knowledge Management, WWW and Electronic Commerce applications development and technologies, and Information Systems Research Methods. The Hypertext Research Laboratory, whose aim is to facilitate the application of hypertext-based technology in academe, business and in the wider community, was founded by him in late 1989.